DOCUMENT RESUME

ED 041 732 SE 008 344

AUTHOR	Harke, Douglas J.
TITLE	Hierarchical Analysis of the Randomized Multiple Choice Format.
INSTITUTION	Purdue Univ., Lafayette, Ind.
PUB DATE	Mar 70
NOTE	21p.; Paper presented at the Annual Meeting of the National Association for Research in Science Teaching (43rd, Minneapolis, Minne., March 5-8, 1970)
EDRS PRICE	EDFS Price MF-$0.25 HC-$1.15
DESCRIPTORS	*Evaluation, *Multiple Choice Tests, *Physics, Scientific Concepts, *Secondary School Science, *Test Validity
IDENTIFIERS	Hierarchical Analysis

ABSTRACT

This paper describes two quantitative methods of hierarchical analysis used to measure the similarity of sequences or hierarchies of steps on two test formats, the Randcmized Multiple Choice (RMS) format, and the free-response problem test format. The investigator first used the Consistency Ratio method to quantitatively assess the transfer relationships for each set of dependency relationships in the hierarchy of RMS items. Consistency ratios for six test problems were calculated. The results indicated that the responses of the subjects as a whole validated most of the dependency relationships within the hierarchies and that in most segments of the hierarchies the students probably were required to use the same problem solving skills and procedures on the RMC and free-response problems. The pattern analysis technique was also used to analyze the responses for the complete hierarchy on a subject by subject basis. Indices of agreement for the six RMC problems were calculated. They revealed the amount of similarity between sequences of physical and mathematical concepts used in solving a problem in the RMC and free-response formats. Bibliography. (LC)

Hierarchical Analysis of the Randomized Multiple Choice Format

Douglas J. Harke*

Purdue University

Many science instructors are not willing to take advantage of the labor saving and high reliability features of machine scored multiple choice tests because of the dichotomous grading of multiple choice items. Questions frequently used on physics tests, for example, require several steps to solve and are usually graded on a continuous scale with the amount of credit awarded being proportional to the degree of correctness of the solution.

The Randomized Multiple Choice (RMC) format was developed to facilitate the awarding of partial credit on machine grading of physics problems which require several steps to solve. This format is basically similar to that suggested by Nedelsky.[4] The problem is stated conventionally and it is suggested to the student to write out a solution to this problem in the space provided below the statement of the problem. Below this space are five or more multiple choice items which correspond to steps in a correct written solution. These multiple choice items have a random order and not the order in which the steps would occur

*Present address is State University College, Geneseo, New York

in a written solution. This is to require the student to organize and use a sequence of physical and mathematical concepts similar to the sequence used in the solution of the same problem in free-response form.

The student completes his written solution and uses it to aid him in answering the RMC items. The student's score on the problem would be the number of multiple choice items answered correctly which should correspond to the number of steps done correctly in the written solution. The student could mark his responses to the RMC items on any standard machine scoreable answer sheet.

The following is an example of an RMC physics test problem.

A car of mass 1500 kg. is travelling at 20 m/sec. when the brakes are suddenly applied. The car takes 20 sec. to come to a complete stop. If the brakes of the car dissipate energy at the rate of 2400 watts to the air while the car is braking, what is the increase in temperature of the brakes? Assume that the heat is absorbed only by the brakes. The brakes (specific heat = 0.10) have a total mass of 40 kg. for the four wheels. 1 kcal. = 4200 j.

1. Before the brakes are applied the car has a kinetic energy of
   A. 60,000 j. B. 600,000 j. C. 15,000 j. *D. 300,000 j.
   E. 30,000 j.

2. If T is the increase of temperature of the brakes, the correct expression for heat absorbed by the brakes is
   A. 40(4200) T kcal. **B. 40(0.10)(4200) T kcal.
   *C. 40(0.10) T kcal. D. $\frac{40\ T}{(0.10)}$ kcal. E. $\frac{40\ T}{4200(0.10)}$ kcal.

---

*correct answer **alternate answer

3. Which one of the following statements is true?
   A. The brakes are capable of dissipating more heat than is produced.
   B. The energy dissipated by the brakes is equal to that absorbed by the brakes.
   *C. The energy dissipated by the brakes is less than that absorbed by the brakes.
   D. The energy dissipated by the brakes is greater than that absorbed by the brakes.

4. The increase in temperature of the brakes is
   A. 63°C. *B. 15°C. C. 630°C. D. 8°C. **E. 0.0036°C.

5. The amount of heat that is retained by the brakes and causes them to heat up is
   *A. 60 kcal. B. 48,000 kcal. C. 15 kcal.
   D. 15,000 kcal. **E. 252,000 kcal.

Some of the RMC items can contain an alternate answer to provide partial credit for a student who made a common error on an initial step but used the correct procedure through the remainder of his solution. The alternate answers are anticipated in advance and are the most common and logical errors that would be made by the students.

When the RMC items are machine-scored, one answer key is submitted for the correct answers and one answer key is submitted for the alternate answers. The scoring machine prints two numbers on the student's card--one which indicates the total number of correct answers marked and one which indicates the total number of alternate answers marked. It should be noted that the RMC format can be graded, without special procedures, by commercially

*correct answer **alternate answer

available machines that are currently used by many educational institutions. Other flexible variations of the multiple choice format have been developed but commerical scoring equipment had to be modified or new computer programs had to be written to grade each test.

The free-response problem test is the most widely used test format in physics and was used as the standard for comparison with the RMC format. Direct comparisons[3] of scores on the RMC and free-response formats were used to determine the similarity of the relative level of performance on the two test formats. These direct comparisons, using correlations, did not reveal whether the students used the same mental processes to solve the same problems on the two test formats. Therefore an indirect approach using hierarchical analysis was employed in comparing sequences of mental skills used on the two test formats.

Some of the objectives of any physics course are to teach the basic concepts in physics and then to give the student experience in using these concepts to solve problems in the physical world. This practice consists of working problems some of which require the application of one concept and some of which require the application of a series of concepts. A problem test in physics is usually given to measure the student's competence in selecting and applying the correct sequence of

physical concepts to solve given problems. For a given problem in free-response form there are very definite sequences in which certain concepts must be applied to solve the problem correctly. These sequences can be logically determined from knowledge of physics.

Each application of a physical concept in a problem solution could be thought of as a step in the problem solution. The logically determined sequences of steps necessary to solve the problem correctly were organized into a hierarchy. Each RMC item corresponded to a step in the written solution to a free-response problem. Therefore the hierarchy of RMC items would be theoretically the same as the hierarchy of steps.

If it can be shown that the sequences in which the physical concepts are used is the same on the RMC format as on the free-response format, then it is quite probable that the same mental processes were used in solving the same problems on the two test formats. Two quantitative methods of hierarchical analysis were used to measure the similarity of sequences or hierarchies of steps on the two test formats.

The Consistency Ratio Method

An analysis similar to the one used by Gagne[2] was employed to quantitatively assess the transfer relationships for each set of dependency relationships in the hierarchy of RMC items.

A dependency relationship was defined as consisting of a given RMC item and the RMC items upon which it was directly dependent. A transfer relationship was defined as the set of dichotomous responses for a given dependency relationship.

From a knowledge of physics it is known that certain steps in a problem must be done correctly before the next step can be done correctly. On this basis it was decided that the following types of transfer relationships validate the hierarchy of RMC items.

1. Mastery of both the upper level item and lower level items.

2. Failure on the upper level item and failure on one or more of the lower level items.

3. Failure on the upper level item but mastery of all of the lower level items.

Transfer relationships of the fourth type where the upper level item is mastered but one or more of the lower level items is failed are contrary to the assumptions underlying the hierarchy and tend to invalidate it.

The first type of transfer relationship would correspond to a situation in a written solution in which the student mastered a certain step and all of the previous steps on which this step was directly dependent. The second type of transfer relationship would correspond to the written solution situation in which the student was unable to master a certain step because he missed

one or more of the previous steps. The third type of transfer relationship would correspond to the written solution situation in which the student was unable to master a certain step but was able to master all of the previous steps. The fourth type of transfer relationship would not occur in a written solution because the student could not master a certain step without being able to master all previous steps.

The number of each type of transfer relationship for each dependency relationship in the hierarchy was tabulated. This procedure was repeated for each problem. Tabulation of the number of cases of each type of transfer relationship in each dependency relationship provided information about performance on each part of the hierarchy. Such an analysis could be useful for modification of the RMC items if the number of cases of the fourth type of transfer relationship rose to an undesireable level.

In order to provide a quantitative measure of validation of each dependency relationship in the hierarchy a consistency ratio was calculated. The consistency ratio was defined as the number of transfer relationships consistent with the hierarchy divided by the total number of transfer relationships. If I, II, III and IV are used to designate the number of cases of each type of transfer relationship, then the consistency ratio can be expressed algebraically as

$$\text{Consistency Ratio} = \frac{\text{I} + \text{II} + \text{III}}{\text{I} + \text{II} + \text{III} + \text{IV}}.$$

A consistency ratio of unity indicated perfect agreement with the hierarchy.

A consistency ratio was calculated for each dependency relationship in each problem. A high consistency ratio would mean that in working an RMC problem the students had to master all of the lower level items connected to an upper level item before they were able to master the upper level item as would be expected on a free-response problem. High consistency ratios on all dependency relationships in the hierarchy would tend to validate the hierarchy. This would suggest that in working the RMC problems the students used very similar or identical sequences to those they would be expected to use if the problems were presented in free-response form.

Results. The results of the transfer relationship analysis are given in Table 1 which shows that the consistency ratios ranged from 0.73 to 0.99 with an average value of 0.88.

The AAAS Science-A Process Approach evaluation committee[1] suggested a consistency ratio equal to or greater than 0.90 for hierarchy validation. This suggested value of 0.90 was apparently for situations corresponding to free-response questions. Guessing on the RMC items could change transfer relationships from the second type to the fourth type and thus reduce the consistency ratio. Therefore the minimum consistency ratio

for each dependency relationship would depend on the number of transfer relationships of the second and fourth type.

Consider, for example, the possible effect of guessing on the #12: #8, #11 dependency relationship in the second problem. There were 92 + 47 or 139 students who were unable to master one or more of the lower level items. By random guessing approximately 35 of these students could have guessed the correct answer to the higher level item. This would reduce to 12 the number of cases contradicting the theoretical hierarchy, or the corrected-for-guessing consistency ratio would be 0.93.

Some of the dependency relationships with a lesser number of transfer relationships of the second and fourth type would have a smaller guessing correction on the consistency ratio. Consider, for example, the dependency relationship #9: #7, #12 in the fourth problem. There were 31 + 45 or 76 students who were unable to master one or more of the lower level items. By random guessing approximately 15 of these students could have guessed the correct answer to the higher level item. This would reduce to 30 the number of cases contradicting the hierarchy, or the corrected-for-guessing consistency ratio would be 0.82.

Some of the transfer relationships of the first type might have been the fourth type had it not been possible for the students to guess the answer to one or more of the lower level

Table 1. Consistency Ratios

Problem	Dependency Relationship	Number of cases with each relationship				Consistency Ratio
		Higher + Lower +	Higher - Lower -	Higher - Lower +	Higher + Lower -	$\frac{(I + II + III)}{(I + II + III + IV)}$
1	#6: #1, #4	23	95	30	25	0.86
1	#1: #5	73	75	23	2	0.99
1	#5: #3, #7	72	70	8	23	0.87
1	#7: #2	74	49	37	13	0.92
1	#4: #2, #3	65	59	37	12	0.94
2	#12: #8, #11	23	92	11	47	0.73
2	#11: #9, #13	24	106	10	33	0.81
3	#4: #6	14	132	14	6	0.96
3	#6: #1, #3	7	130	8	21	0.87
3	#3: #2, #5	59	57	21	29	0.83

Table 1 (cont'd.)

Problem	Dependency Relationship	Number of cases with each relationship				Consistency Ratio
		Higher + Lower +	Higher - Lower -	Higher - Lower +	Higher + Lower -	$\frac{(I + II + III)}{(I + II + III + IV)}$
4	#10: #9, #11	60	52	15	39	0.77
4	#11: #7, #12	57	47	33	29	0.83
4	#9: #7, #12	76	31	14	45	0.73
5	#2: #4, #7	43	82	15	16	0.90
5	#7: #5	76	54	23	3	0.98
5	#5: #3, #6	85	49	8	14	0.91
6	#8: #11	51	65	24	16	0.90
6	#11: #9, #12	61	45	36	14	0.91
6	#12: #10	108	9	33	6	0.96

items. However there was less possibility of this happening than the possibility of guessing converting the second type to the fourth type of transfer relationship because the p-values of the lower level items were greater than the p-values of the upper level item. The frequency of the first type of transfer relationship, which appeared low rather than inflated, provided further evidence that guessing did not change a great number of relationships from the fourth type to the first type.

The effect of guessing could have changed some of the transfer relationships of the fourth type to the first type and thus increased the consistency ratio. The guessing correction which could be applied to the second and fourth types of transfer relationship was probably an undercorrection. Therefore the amount of undercorrection would probably be compensated for by the tendency of guessing to slightly reduce the frequency of the fourth type of transfer relationship.

The value 0.85 was selected as the minimum consistency ratio necessary to validate a hierarchy in which the questions are in multiple choice form. This was probably a fairly safe choice because it has been shown earlier that guessing could reduce the consistency ratio by as much as 0.2. In most cases the reduction was probably closer to 0.1.

Four of the six RMC problems used had average consistency

ratios above the value 0.85. One of the consistency ratios in the third problem was below 0.85 which indicated a possible weak link in the hierarchy. All of the consistency ratios in the second and fourth problems were below 0.85.

The overall results indicated that the responses of the group as a whole validated most of the dependency relationships within the hierarchies. The results were conclusive on the three problems which had all of their consistency ratios above 0.85. The average consistency ratio on the third problem, which had one consistency ratio below 0.85, was above 0.85. Therefore the results on the third problem probably also validated the proposed hierarchy. Two of the problems each had several low consistency ratios because it was found later that a two dimensional hierarchy was probably not an adequate representation of the interrelationships of the correct and alternate answers.

If the average consistency ratio was an appropriate measure of the validity of the total hierarchy, then the experimental results indicated that the RMC item responses validated the hierarchies which were constructed on the basis of the interrelationships of the steps in a written solution. The high consistency ratios specifically showed that the relationships between certain RMC items were apparently the same as the relationships between certain steps in the written

solution. This indicated that at least in most segments of the hierarchies the students probably were required to use the same problem solving skills and procedures on the RMC and free-response problems.

Pattern Analysis

The consistency ratio method considered group responses as a whole on the complete hierarchy and group responses on segments of the hierarchy. The pattern analysis technique developed by Rimoldi and Grib[5] was used to analyze the responses for the complete hierarchy on a subject by subject basis.

Rimoldi and Grib developed a versatile method of pattern analysis to compare bivariate patterns in which a number of subjects respond to a number of items. The responses to the items may be dichotomous or may be assigned other numerical values. If the test vector of the subjects are combined, the matrix formed will have rows corresponding to subjects and columns corresponding to items. This observed matrix can be compared to an expected matrix using the technique of Grib and Rimoldi. The expected matrix can be derived from any model. The only restriction on the expected matrix is that if the patterns are being compared across subjects then the subject's total score on the expected pattern must equal his total score on the observed pattern. A similar restriction

would be placed on the columns if the patterns were compared across items.

Once the expected pattern has been generated, weights are calculated for each cell of the expected pattern. The weights a_{ij} for the cells containing ones are given by

$$a_{ij} = \frac{R_i C_j}{\Sigma R_i}$$

R_i is equal to the number of ones in the ith row.

C_j is equal to the number of ones in the jth column.

The weights $\bar{a}_{ij}$ for the cells containing zeros are given by

$$\bar{a}_{ij} = \frac{\bar{R}_i \bar{C}_j}{\Sigma R_i}$$

$\bar{R}_i$ is equal to the number of zeros in the ith row.

$\bar{C}_j$ is equal to the number of zeros in the jth column.

This method of assigning weights makes no a priori assumptions as to the pattern expected.

The amount of agreement or correlation between two patterns is expressed by the index of agreement I_a. When patterns are compared across subjects, this index of agreement can be written as

$$I_a = \frac{A_t - {}_mA_t}{T - {}_mA_t}.$$

T is the sum of all a_{ij} and $\bar{a}_{ij}$.

A_t is the sum of a_{ij} and $\bar{a}_{ij}$ of the cells that are the same in the expected and observed patterns.

${}_mA_t$ is the sum of a_{ij} and $\bar{a}_{ij}$ corresponding to minimum possible agreement.

Certain patterns like those containing mostly zeros or mostly ones have less permutations than a pattern containing an equal number of ones and zeros. The possibility of disagreement increases with the number of permutations. The term ${}_mA_t$ therefore reduces the power of the patterns containing mostly ones or mostly zeros.

The index of agreement varies between unity for perfect agreement and zero for no agreement. Grib and Rimoldi compared values of I_a to the coefficients obtained using other pattern analyses and found that I_a gives conservative values. No significance test has yet been developed for the index of agreement.

Hierarchical procedures could provide a method of investigating the responses to interconnected test items to determine whether the sequence of responses indicated or validated certain patterns.

Only a finite number of sequences of dichotomous elements (1 for correct and 0 for incorrect) were possible according to the relationships within the hierarchy of RMC items. The students' response patterns on the RMC items within a problem

were examined to determine whether they corresponded to one of the expected patterns. If a student's response pattern did not correspond to one of the expected patterns, then the expected pattern most closely resembling and having the same number of each type of element as the observed pattern was chosen as the expected pattern for that student. These expected and observed response patterns were compared using the Rimoldi and Grib procedure. This procedure was carried out and the index of agreement was computed for each of the six RMC problems.

The index of agreement was interpreted as the amount of similarity between the sequences of physical concepts used in solving a free-response problem and the same problem presented in RMC format.

The indices of agreement for the six RMC problems used are given in Table 2.

Table 2. Indices of Agreement

Problem	1	2	3	4	5	6
Index of Agreement	0.789	0.640	0.812	0.705	0.939	0.819

The indices of agreement revealed the amount of similarity between sequences of physical and mathematical concepts used in solving a problem in the RMC and free-response formats.

The low indices of agreement on the second and fourth problem may have been the result of the inadequacies of the hierarchical model chosen.

The index of agreement could best be thought of as a correlation coefficient because the group was compared subject by subject on the expected and observed response patterns. Because this method reduced the contributions made by patterns with few possible permutations, the index of agreement probably was an underestimate of the correlation of the expected and observed response patterns.

Ideally the index of agreement should have been unity but the error introduced by guessing reduced the maximum index of agreement attainable. There was no way of quantitatively assessing how much the index of agreement was reduced by guessing.

The relatively large indices of agreement obtained indicated that the individual student's response patterns to the RMC items closely resembled the response patterns expected on the basis of the interrelationships of steps in a written solution. This further indicated that the students, for the total problem, used the same sequences of mathematical and physical concepts on the RMC format that they would have used if the problems would have been presented in free-response form. Therefore it is very probable that in working the total problem

the students as individuals used the same problem solving skills and procedures on the RMC and free-response problems.

Conclusions

The results of the hierarchical analyses indicates that the students in solving the RMC problems used sequences of mathematical and physical concepts similar to the sequences they would have used if the same problems had been presented in free-response form. Since this indicated that the students proceeded through the steps of the RMC problem solution in the same order as they would in a written solution, it could be inferred that the students probably used the same mental processes on the two test formats. This argument could be further strengthened by considering the following example--in a particular free-response problem solution the student might be required to apply two different physical concepts and then combine the results of these applications to arrive at the final answer. In this example, the student, in addition to knowledge of the physical concepts, was required to use cognitive skills similar to those of application and synthesis. Analysis of the students' response patterns to the RMC items indicated that they usually correctly answered the RMC items corresponding to application of the concepts before they could correctly answer the RMC item which required the synthesis of

the two previous items. Therefore it is highly likely that the RMC problems required the same or very similar cognitive processes of problem solving skills as the free-response problems.

References

1. AAAS Commission on Science Education. Science-A Process Approach: An Evaluation and its Application, Second Report. American Association for the Advancement of Science, 1968.

2. Gagne, R. M. The Acquisition of Knowledge. Psychological Review, 1962, 69, 355-365.

3. Harke, D. J. Evaluation of the Randomized Multiple Choice Format. (Doctoral dissertation, Purdue University) Ann Arbor, Mich.: University Microfilms, 1969.

4. Nedelsky, L. Evaluation of Essays by Objective Tests. Journal of General Education, 1953, 7, 209.

5. Rimoldi, H. J. A. & Grib, T. F. Pattern Analysis. British Journal of Statistical Psychology, 1960, 13, 137-149.